agriculture

MDPI

Article

Counting Dense Leaves under Natural Environments via an Improved Deep-Learning-Based Object Detection Algorithm

Shenglian Lu [1], Zhen Song [1], Wenkang Chen [1], Tingting Qian [2,*], Yingyu Zhang [2], Ming Chen [1] and Guo Li [1]

[1] Guangxi Key Lab of Multisource Information Mining & Security, College of Computer Science & Engineering, Guangxi Normal University, Guilin 541004, China; lsl@gxnu.edu.cn (S.L.); songzhen1992@gmail.com (Z.S.); cwk1031645988@gmail.com (W.C.); mingchen@gxnu.edu.cn (M.C.); liguo@gxnu.edu.cn (G.L.)

[2] Agricultural Information Institutes of Science and Technology, Shanghai Academy of Agriculture Sciences, Shanghai 201403, China; yingyu.zhang@saas.sh.cn

* Correspondence: qiantingting@saas.sh.cn; Tel.: +86-150-0075-3513

Abstract: The leaf is the organ that is crucial for photosynthesis and the production of nutrients in plants; as such, the number of leaves is one of the key indicators with which to describe the development and growth of a canopy. The irregular shape and distribution of the blades, as well as the effect of natural light, make the segmentation and detection process of the blades difficult. The inaccurate acquisition of plant phenotypic parameters may affect the subsequent judgment of crop growth status and crop yield. To address the challenge in counting dense and overlapped plant leaves under natural environments, we proposed an improved deep-learning-based object detection algorithm by merging a space-to-depth module, a Convolutional Block Attention Module (CBAM) and Atrous Spatial Pyramid Pooling (ASPP) into the network, and applying the $smooth_{L1}$ function to improve the loss function of object prediction. We evaluated our method on images of five different plant species collected under indoor and outdoor environments. The experimental results demonstrated that our algorithm which counts dense leaves improved average detection accuracy of 85% to 96%. Our algorithm also showed better performance in both detection accuracy and time consumption compared to other state-of-the-art object detection algorithms.

Keywords: deep learning; plant phenotyping; leaf detection; object recognition

Citation: Lu, S.; Song, Z.; Chen, W.; Qian, T.; Zhang, Y.; Chen, M.; Li, G. Counting Dense Leaves under Natural Environments via an Improved Deep-Learning-Based Object Detection Algorithm. *Agriculture* **2021**, *11*, 1003. https://doi.org/10.3390/agriculture11101003

Academic Editor: Domenico Pignone

Received: 7 September 2021
Accepted: 29 September 2021
Published: 14 October 2021

1. Introduction

The phenotype of plant refers to all observable characteristics of a plant, including its physical morphology as well as biochemical and physiological properties. These characteristics are influenced by both genetic code expression and environment, which can consequently change during the growth of plants. Thus, rapid and accurate techniques of plant phenotype detection play an important role in studying genetic and environmental synergistic effects on crop growth, and have attracted much attention in agronomy research [1]. The benefit of advanced computer vision, robotics and artificial intelligence technologies, being the efficient detection of plant phenotypic traits, has also implied progressive developments in recent years [2,3]. Early measurements of these plant architectural features were often carried out manually. However, the manual collection of architectural phenotypic traits can be time-consuming and labor-intensive. Bao et al. [4] found that, depending on the growth stage, 32 to 64 man-hours of manual measurements were necessary to study 11 traits in 18 sorghum plots. In comparison, these plots could be imaged within three minutes using a high-throughput phenotyping (HTP) platform. Therefore, numerous studies on automated plant phenotyping in greenhouses, fields and controlled laboratory conditions have been carried out in the last decade. Automated plant phenotyping usually focuses on parameters that are related to architectural indicators (height, leaf number, leaf angle, etc.). Especially, the number of leaves of a plant is considered as one of the key phenotypic metrics related to its development rate, flowering

time or yield potential. Therefore, great efforts have been devoted to the identification and detection of leaves from digital images.

In early works on leaf detection, image analysis techniques based on color or shape features were usually used for the identification or segmentation of leaves [5]. For example, based on the HSI color space model (which represents colors with three components: hue, saturation and intensity), Tang et al. [6] used a watershed segmentation method to extract the leaf region from plant images with a complex background. Yin et al. [7] matched the existing extracted leaf templates with hidden data to segment and track the Arabidopsis leaf. Another study used an active contour method to segment and track the leaves of Arabidopsis plants [8]. As proposed by Dellen et al. [9], the graph-based method can also be used for leaf tracking. Grand-Brochier et al. [10] reported a comparative study on tree leaf extraction from natural images. Cerutti et al. [11] developed a parametric active polygon model for leaf segmentation and shape estimation. However, these methods are dedicated to detecting separate leaves, and often fail when handling overlapped leaves. Additional works have thus been carried out to address the issue of segmenting overlapped leaves. Pape and Klukas [12] used a 3D histogram of the L*a*b* color space of plant images for the supervised segmentation of foreground/background, in which a distance map, skeleton and equivalent graph representation are used to find the separate leaves. Scharr et al. [13] analyzed the use of the region growing method and its drawbacks for segmenting leaves over the superpixels extracted from the L*a*b* color map. Vukadinovic and Polder [14] proposed a method of combining a supervised classification and a simple artificial neural network to segment the plant regions, and then using watershed transformation to identify individual leaves. Nevertheless, this approach uses ground truth images to mask the plant and background pixels, and therefore is not suitable for automatic processing. Giuffrida et al. [15] used log-polar representation and global descriptors to estimate the number of leaves in a plant. Apart from the above, some efforts have also been made to count objects via density estimation by per-pixel ridge and random forest regression [16,17]. This method works by learning a mapping, F: X Y, between local image features, X, and object density, Y, which then allows the derivation of an estimated object density map for unseen images, where regressors are used to infer local densities. However, these approaches in handling leaf occlusions can be influenced by large-scale variability, moreover, these methods would also suffer from significant limitations, such as requiring specific light conditions or only being suitable for specific plants [18].

Recently, artificial intelligence methods such as convolutional neural networks (CNNs) have made important progress in plant phenotyping, and have brought wide applications in plant classification, fruit detection and leaf segmentation [19]. The leaf counting can be addressed by using semantic segmentation or instance segmentation from a deep learning perspective. Compared to conventional machine learning methods, deep neural networks showed better accuracy and faster processing efficiency [20]. Some solutions used recurrent neural networks to segment leaves and count fruits. For instance, Romera-Paredes and Torr [21] developed an end-to-end model of recurrent instance segmentation by combining convolutional long short-term memory (LSTM) [22] and spatial inhibition modules, thereby segmenting one leaf at a time by keeping track of the spatial information within each image. Ren and Zemel [23] applied visual attention to jointly compute instance segmentation with a counting function. This method obtains sequential attention by creating a temporal chain via a LSTM cell which outputs one instance at a time, in which non-maximal suppression was used to solve heavily occluded scenes. Aich and Stavness [24] used deep convolutional and deconvolutional network (DCDN) to segment the rosette plant region and to count the rosette plant leaves. Giuffrida et al. [25] used ResNet50 architecture to count the rosette plant leaves from multimodal 2D images. Although their network was pretrained on only the *Arabidopsis* images in the CVPPP dataset, the network runs well on other datasets (e.g., tobacco). Recently, Kumar and Domnic [26] used the circular Hough transform and DCNN models to count the leaves from segmented plant regions. Their detection results on leaves of rosette plants reached 0.96, 0.94 and 0.95 on average precision, recall and F_1 score,

respectively. In general, LSTM is a type of temporal recursive neural network (RNN) and is more suitable for time series data, while DCDN works by using a deconvolutional network for initial segmentation and a convolutional network for instance counting; therefore, DCDN is expected to achieve better performance in object detecting and counting.

Most existing methods for detecting and counting overlapped leaves were dedicated to leaves of rosette-shaped plants such as *Arabidopsis*, which has a relatively simple morphological structure and less overlapping on its canopy. In this work, the state-of-the-art deep learning-based instance segmentation detector—CenterNet—is applied and improved, to address the problem of counting dense, overlapped and size-changing leaves in different plant canopies under natural environments. The main objectives of this study were:

(1) to architecturally improve the current CenterNet network (i.e., improved CenterNet) with the specific focus on overlapped leaves in canopy images captured under natural light conditions;
(2) to verify the performance of our improved CenterNet with other commonly used deep neural networks, including Mask R-CNN, Faster R-CNN, YOLOv4 and original CenterNet, as well as a commonly used traditional machine learning method;
(3) to validate the value and significance of our improved CenterNet by using several plant species with different shapes and sizes in leaves.

2. Materials and Methods

CenterNet [27] is a key point-based one-stage object detection algorithm, with a simple architecture and excellent detection performance. CenterNet uses key point estimation to find center points and regresses to all other object properties. It adopts a fully convolutional upsampling DLA-34 backbone network, with skip connections between lower layers to the output augmented by deformable convolution (see the network structure in Figure 1). According to benchmark tests based on the COCO dataset [28], the DLA-34 network achieved a detection speed of 52 fps with a recognition accuracy of 37.4%, while the accuracy of CenterNet is 4% higher than that of YOLOv3 at the same detection speed. YOLO (You Only Look Once) is a state-of-the-art object detection and localization network [29], while YOLOv3 is the third version of YOLO.

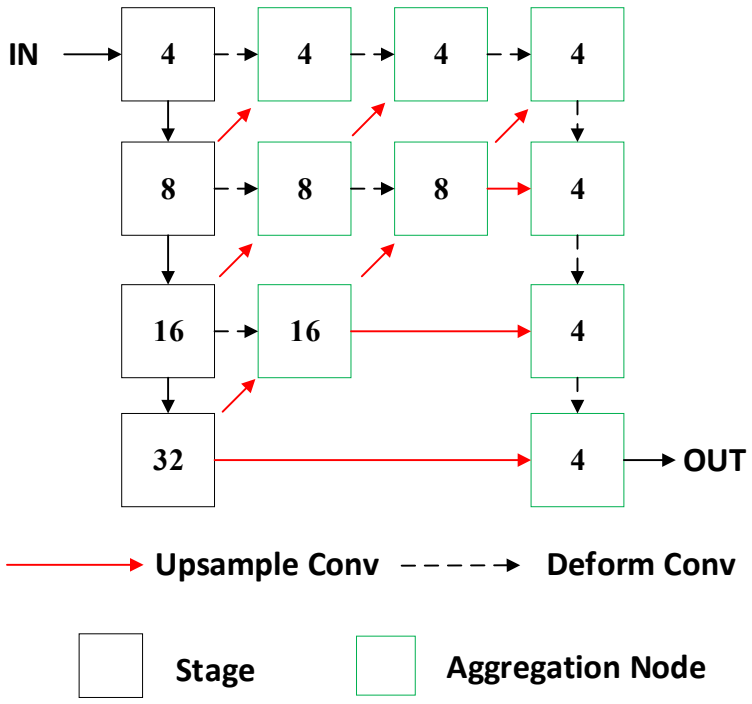

Figure 1. Network structure of DLA-34 with fully convolutional upsampling.

2.1. Motivation in Improving CenterNet Algorithm

As discussed above, the number of leaves is usually one of the key parameters in plant phenotyping. The key of leaf segmentation and detection is to accurately extract the position, size, shape and other information of each leaf. However, dense leaves are often overlapping each other and usually under complex light conditions. Thus, CenterNet algorithm would still meet challenges posed by leaf detection.

To overcome these challenges, several modifications were made to the original CenterNet algorithm:

(1) Since leaves have significantly different sizes at different growth stages, a space-to-depth module was used to convert the input image into feature maps with different resolutions. In this way, the detection performance for small leaves could be improved by high-resolution feature maps;

(2) For accurate extraction of edge information of overlapped leaves, an attention mechanism CBAM was used behind each resolution feature map. This allowed the network to focus on the key information that distinguishes the edge features of each leaf.

(3) To address the influence of different factors and the effective combination of different features, atrous convolutions and atrous spatial pyramid pooling (ASPP) modules were used to extract the image receptive field features at different scales. By combining the feature maps detected at different scales, the network could store more image information to facilitate the detection of plant leaves.

Details of the above modifications will be given in the following.

2.2. The Improved Network Structure

As shown in Figure 2, the network structure of the improved CenterNet was composed of a space-to-depth module, a CBAM [30] attention mechanism module, an ASPP module [31], a deep feature fusion network DLA-34, and a feature fusion and level chain module.

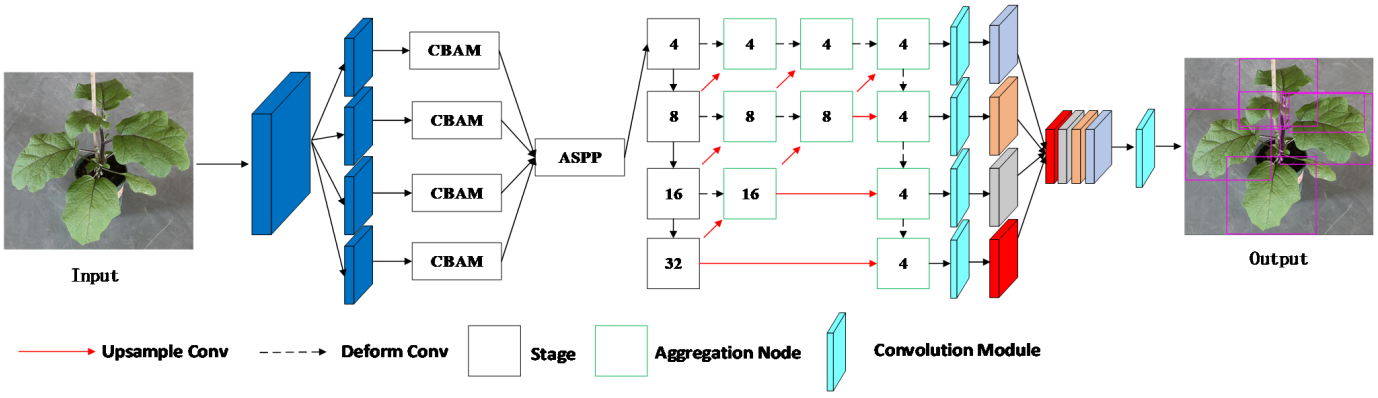

Figure 2. The improved network structure of CenterNet for detecting dense and overlapped leaves in plant canopy.

The backbone network was sampled at different spatial resolutions to output feature maps. Each feature map had a specific spatial resolution and could be viewed as a stage. At each stage, the residual block method [32] was used to generate feature maps with different spatial resolutions, and obtain leaf features with different spatial resolutions. Then, the attention mechanism was introduced to allow the network to focus on the edge information of each leaf and correctly identify each leaf region. Finally, these feature maps were connected, and the leaf was detected based on the cascaded feature maps with 1/4 input resolution.

2.3. Space-to-Depth Module

If the ASPP is only used in downsampling processes, a feature degradation issue is likely to be raised, which is unfavorable for obtaining object information. Meanwhile, to segment and detect overlapped leaves with different shapes and sizes, the edge details of the leaves must be properly determined. It requires not only a high resolution in feature maps, but also good detection performance for feature maps under different scales. Based on the above analysis and the previous work by Li et al. [33], a space-to-depth module was used to extract detailed features of leaves at smaller image resolution. It was accomplished by dividing the input data equally into four blocks based on the height and width of the image; then, stacking these four blocks to generate a new block with a smaller size but deeper depth.

2.4. Attention Mechanism

The attention mechanism has been used widely in natural language processing, image recognition and speech processing [34]. Considering that the size/shape of a leaf may be significantly different at different growth stages or in different plant species, the attention mechanism can enhance object features while suppress non-object features of leaves, highlights object information and understates background information, thus reaching better detection accuracy.

In this study, we combined the CBAM into the improved CenterNet. CBAM combines both channel sub-module and spatial attention sub-module. The channel sub-module squeezes the spatial dimensions of the feature map firstly, thus obtaining a one-dimensional vector for further operations. It involves both average-pooling and max-pooling, which yields two one-dimensional vectors. Overall, the channel attention is calculated as:

$$M_C(F) = \sigma(MLP(AvgPool(F)) + MLP(MaxPool(F))) = \sigma(W_1(W_0(F^c_{avg})) + W_1(W_0(F^c_{\max}))) \quad (1)$$

where, σ is the sigmoid activation function, F denotes the input feature map, F^c_{avg} and $F^c_{\max}$ respectively denote average-pooled and max-pooled feature map, W_0 and W_1 are two layers of parameters in the multilayer perceptron model.

In the spatial sub-module, the channel dimension is squeezed by applying average-pooling and max-pooling operations. The max/average-pooling extracts the max/average values along the channel, with many extraction of height $\times$ width. After the extraction, the extracted feature maps (with 1 channel) are combined to generate a new feature map (with 2 channels). This is accomplished by:

$$M_S(F) = \sigma(f^{7\times7}([AvgPool(F); MaxPool(F)])) = \sigma(f^{7\times7}([F^s_{avg}; F^s_{\max}])) \quad (2)$$

Here the convolution layer uses a 7×7 kernel, as it shows better performance than a 3×3 kernel. The CBAM module is a general and lightweight module and can be readily integrated into the convolution module of any network, thus enabling end-to-end training. In this study, after the image was converted with the space-to-depth module, the CBAM module is used as convolution module with different depths. The integrated attention mechanism enables the network to extract more key information from the image, and improves the detection accuracy of overlapping and small objects.

2.5. Atrous Spatial Pyramid Pooling

Atrous spatial pyramid pooling (ASPP) is a module for sampling a given input image at multiple rates using multiple parallel atrous convolutional layers. In general, it captures objects and useful image context at multiple scales [35]. Atrous convolutional with different sampling rates can effectively extract image receptive field features at different scales, while the ASPP structure further enhances the image segmentation effect by combining multi-scale image information with different-scale cavity convolutions, thereby achieving more accurate and efficient classification. The ASPP module is composed of two main parts (see Figure 3):

- 1×1 and 3×3 convolutions, with dilation rates of 6, 12, and 18;
- image-level features.

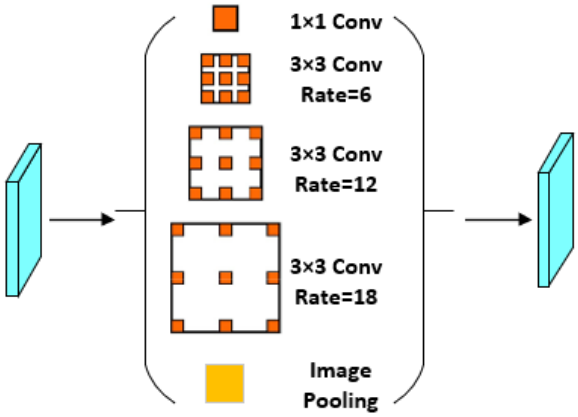

Figure 3. Illustration of Atrous Spatial Pyramid Pooling.

Convolution is given by:

$$Kd = K + (K-1) \times (d-1) \tag{3}$$

where, K is initial kernel size, d is the atrous rate. The number of sampling points in ASPP is 3, with a receptive field $K = 13$.

Overlapped leaves usually have irregular shapes, information from different receptive fields is valuable for plant leaf detection. Meanwhile, significant amounts of feature information can be lost during convolution and pooling, resulting in a decrease in detection accuracy. In this study, we combined ASPP module (with dilation rates of 1, 2, 3, and 6) with a deep feature network to overcome the shortcomings of overlapping or irregular object detection under different receptive fields. Different scaled images with attention mechanism were integrated with the ASPP module which improved the ability to segment overlapping leaves. The fused feature map was then inputted into the deep feature network, which produced a feature map with rich information between different deep networks.

2.6. Modification on Loss Function

The loss function of CenterNet includes three parts: the loss function of the heat map, the object size prediction loss function, and the center point offset loss function. Note that correct frame selection of dense leaves is the key of accurate leaf detection. Therefore, to improve the performance of leaf detection, we modified the object size prediction loss function in the CenterNet loss function.

In our leaf datasets, the density of leaves is relatively high. In this way, two leaves in an image can share one feature point after being down-sampled by a factor of 16, thus are difficult to distinguish. A small leaf object can reduce to 1 pixel or even diminish after being down-sampled by a factor of 32, making feature extraction rather difficult. To accurately predict leaf size and to increase the robustness for outliers, we used the $smooth_{L1}$ loss function [36] for object size learning and training. Compared with the L2 loss function, the $smooth_{L1}$ loss function is less sensitive to outliers and anomalies, thus has better gradient control and convergency, ensuring accurate detection on leave edges with different shapes. The modified loss function for object size is given by:

$$L_{size} = \frac{1}{N} \sum_{k=1}^{n} smooth_{L1}(\hat{S}_{p_k} - S_k) \tag{4}$$

Here, N denotes the number of key points in the image, $\hat{S}_{p_k}$ is the predicted object size, S_k is the actual size the object.

During the model training, the Focal loss function [37] is adopted for logistic regression. While the center point offset value loss function is defined as:

$$L_{off} = \frac{1}{N}\sum_{p}\left|\hat{O}_p - \left(\frac{P}{R} - \hat{P}\right)\right| \quad (5)$$

where, $\hat{O}_p$ is the predicted offset, P denotes the center point coordinates of the leaf in the image, R is the scaling factor, $\hat{P}$ is the rounded integer coordinates of the center point after image scaling.

3. Experiments and Discussion

3.1. Data Acquisition and Organization

3.1.1. Data Acquisition

All images in this work were collected at Shanghai Academy of Agricultural Sciences. We obtained images of cucumbers in a plastic greenhouse by placing a high-resolution camera directly above the canopy (Figure 4). The images were taken hourly from 6:00 to 18:00 every day, from seedling to flowering stage. A total of 300 images were collected, including images under different weather (sunny and cloudy) and light (direct light and back light) conditions. Images of eggplant, tomatoes, pennywort, and orchid grass were also collected in a glass greenhouse by using an EOS 5D Mark III camera (Canon, Oita Prefecture, Japan). For each type of crop, we prepared 5–15 pots of plants and took 30 images from different angles. The resolution of each image is 3000 × 4000 pixels.

(**a**) (**b**)

Figure 4. Image acquisition scene: (**a**) in glass greenhouse; (**b**) in plastic greenhouse.

3.1.2. Image Labeling and Data Augmentation

The LabelImg package were used to label plant images. Both the position and the type of each leaf were marked during the image labeling. In this study, we considered two situations, i.e., with over/under 50% of leaf area being overlapped (by other leaves), respectively. Leaves with overlapped rate over 80% were not labeled. Augmentor image enhancement library was used to augment our leaf dataset. According to the different interference factors involved in our experimental environments, the augmentation methods considered here include rotating the original image (Figure 5a,c), adjusting the brightness of the image (Figure 5b), and increasing the noise (Figure 5d). As a result, the original leaf dataset was expanded by 10 times.

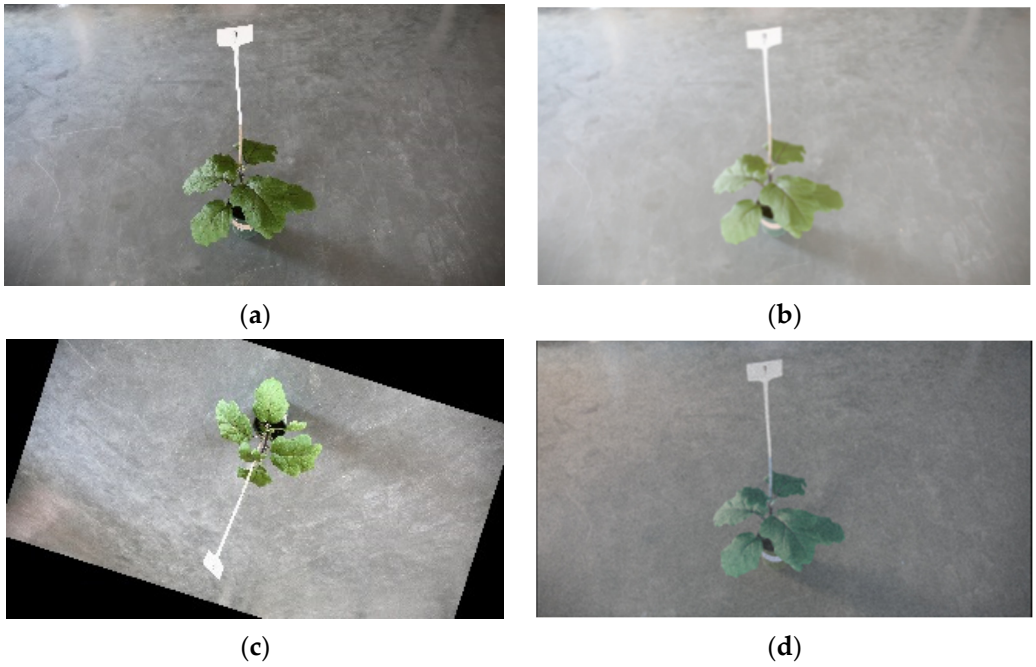

Figure 5. Image augmentation by: (**a**) original image; (**b**) adjust brightness; (**c**) rotation; (**d**) increase noise.

The structure of the dataset and directory was organized similar to the COCO dataset. The dataset was divided into training set, test set, and validation set; represent 80%, 10%, and 10% of the entre dataset, respectively. The whole processing flow chart of our proposed method is shown in Figure 6.

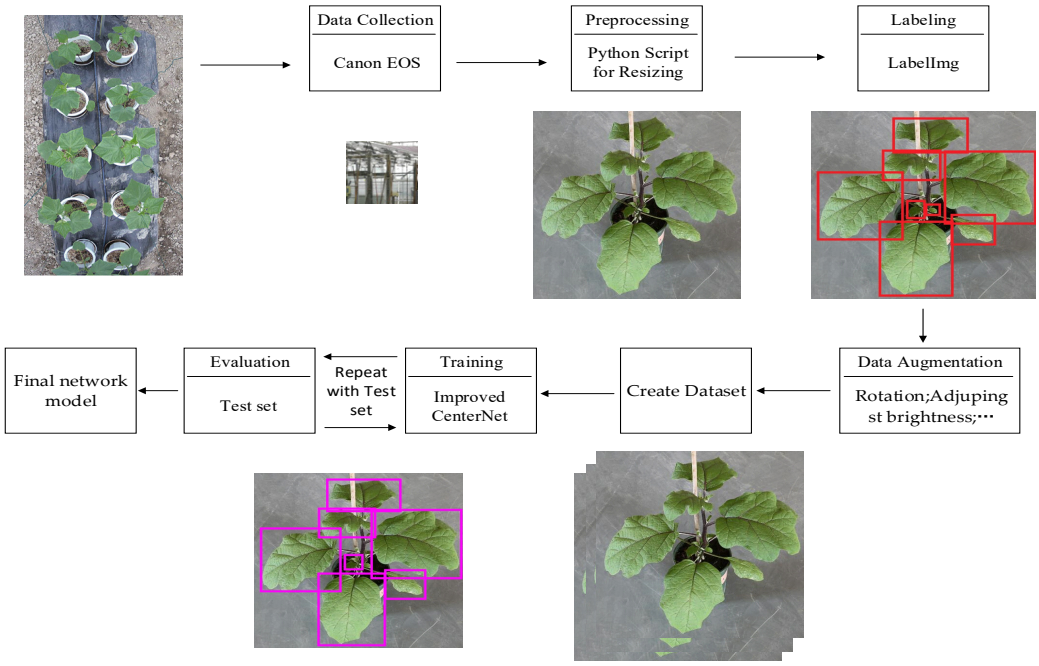

Figure 6. Overall processing flow chart.

3.2. Training and Testing

3.2.1. Implementation Protocol and Evaluation Parameters

We used the Detectron2 deep-learning platform based on programming language Python. All experiments were performed on an Intel Core i7-9850H CPU and a Quadro P4000 GPU, with CUDA 10.1 and CUDNN 7.6.5. Three state-of-the-art deep learning-based object detection methods (Mask R-CNN [38], Faster R-CNN [39], YOLOv4 [40]) and a traditional machine learning method (ExG + SVM [41]) were used for comparison.

The relevant parameters for evaluating the effectiveness of the model include: precision, recall, model size, and detection time [42].

3.2.2. Training

Two different training frameworks were used for these detection models, where Faster R-CNN, Mask R-CNN, the original CenterNet, and the improved CenterNet were trained with PyTorch, while YOLOv4 was trained with its own training framework. All these models were pre-trained by using the leaf dataset we collected and organized. Before training, we also augmented the data with random flips, random scaling (between 0.6 and 1.3), cropping, as well as color dithering, and optimized the overall goal with Adam's algorithm [43].

3.3. Comparison with the State-of-the-Art Methods

The improved method were evaluated with the other state-of-the-art methods mentioned above by using images of cucumber leaves. P-R curves of the six object detection methods are shown in Figure 7. P-R curves in Figure 7 showed that the improved CenterNet model has better performance than the other five methods. We compared the detection performance of dense leaves on early and late stage of cucumber. Table 1 shows the relevant parameters of the test results of each method. The detection accuracy of the improved CenterNet is 96% for early cucumber leaves, with a detection speed of 0.29 s for each image.

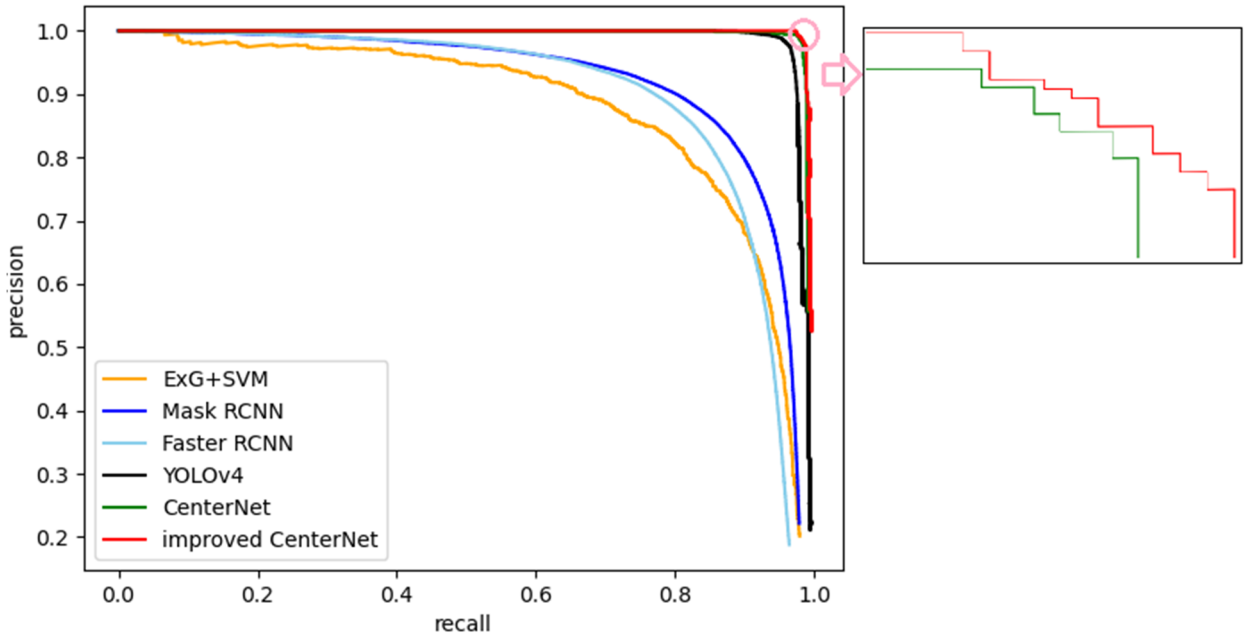

Figure 7. P-R curves for six detection detectors.

Even for dense late-stage leaves, the accuracy of improved CenterNet still reaches 92%. Despite the model size of improved CenterNet is larger than that of the original CenterNet, the accuracy and recall of the improved CenterNet is the highest among the six models. Figure 8 gives a visual leaf detected result on late-stage cucumber.

Table 1. Relevant parameters of six object detection methods in detecting cucumber leaves.

Model	Number of Training Images	Growth Stage	Model Size/M	Detection Time (s)	Precision	Recall	Training Time (h)
ExG + SVM	300	Early	—	0.76	0.78	0.74	16
Mask R-CNN	300	Early	251	3.13	0.85	0.80	23
Faster R-CNN	300	Early	540	3.52	0.82	0.79	21
YOLOv4	300	Early	250	0.46	0.92	0.86	40
Original CenterNet	300	Early	62	0.24	0.94	0.87	21
Improved CenterNet	300	Early	83	0.29	**0.96**	**0.90**	**22**
ExG + SVM	300	Late	—	0.76	0.71	0.68	16
Mask R-CNN	300	Late	251	3.13	0.82	0.78	23
Faster R-CNN	300	Late	540	3.52	0.8	0.76	21
YOLOv4	300	Late	250	0.46	0.88	0.84	40
Original CenterNet	300	Late	62	0.24	0.9	0.86	21
Improved CenterNet	300	Late	83	0.29	**0.92**	**0.89**	**22**

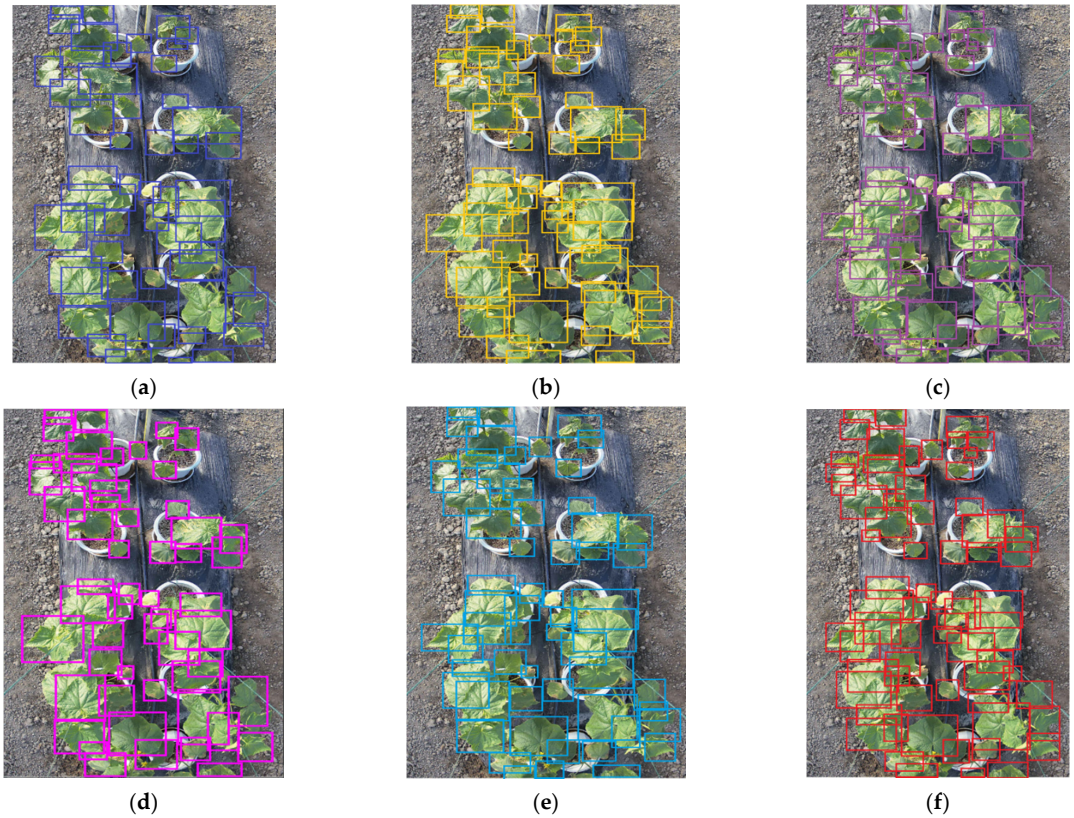

Figure 8. Visual effects of six leaf detection methods on the late stage of cucumber: (**a**) ExG + SVM; (**b**) Mask R-CNN; (**c**) Faster R-CNN; (**d**) YOLOv4; (**e**) original CenterNet; (**f**) the improved CenterNet.

3.4. Detection Performance for Leaves with Different Shape and Silhouette

To compare the detection performance for leaves with different shape and silhouette, images of five different plant species were collected, including cucumber, tomato, eggplant, orchid grass, and pennywort (30 images for each plant). Images of eggplants and tomato were collected outdoor. Data augmentation method was also used to expand the leaf image number to 300 per plant for training.

As can be seen from Figure 9, the improved CenterNet shows the best overall performance for these test plants, with the detection accuracy for Pennywort of 96%, which is about 3% higher than the detection accuracy of YOLOV4. Figure 10 gives the visual detected results on the leaves of six plant species mentioned above. From Figure 10 we can

see that the orchid grass has slender leaves, the pennywort leaves are generally rounded, the cucumber leaves are small at early stage, but become larger and overlapped at late stage, with some leaves being only partially visible to the camera. The improved CenterNet presents a good detection performance with high detection accuracy and speed. Even in the case for late stage leaves with significant overlapping, the detection accuracy of the improved CenterNet is still above 87%.

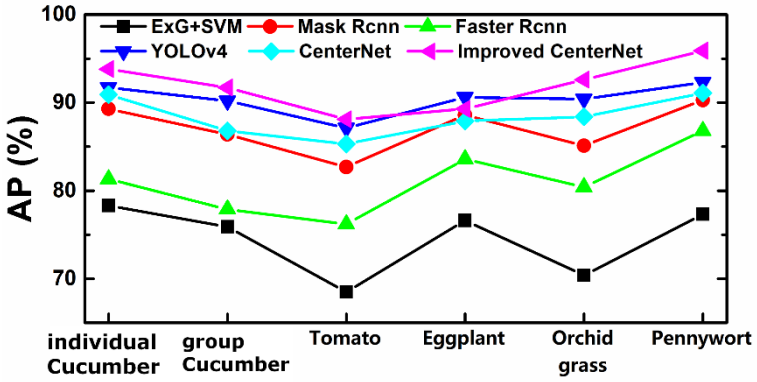

Figure 9. Average detection accuracy for leaves of different plants.

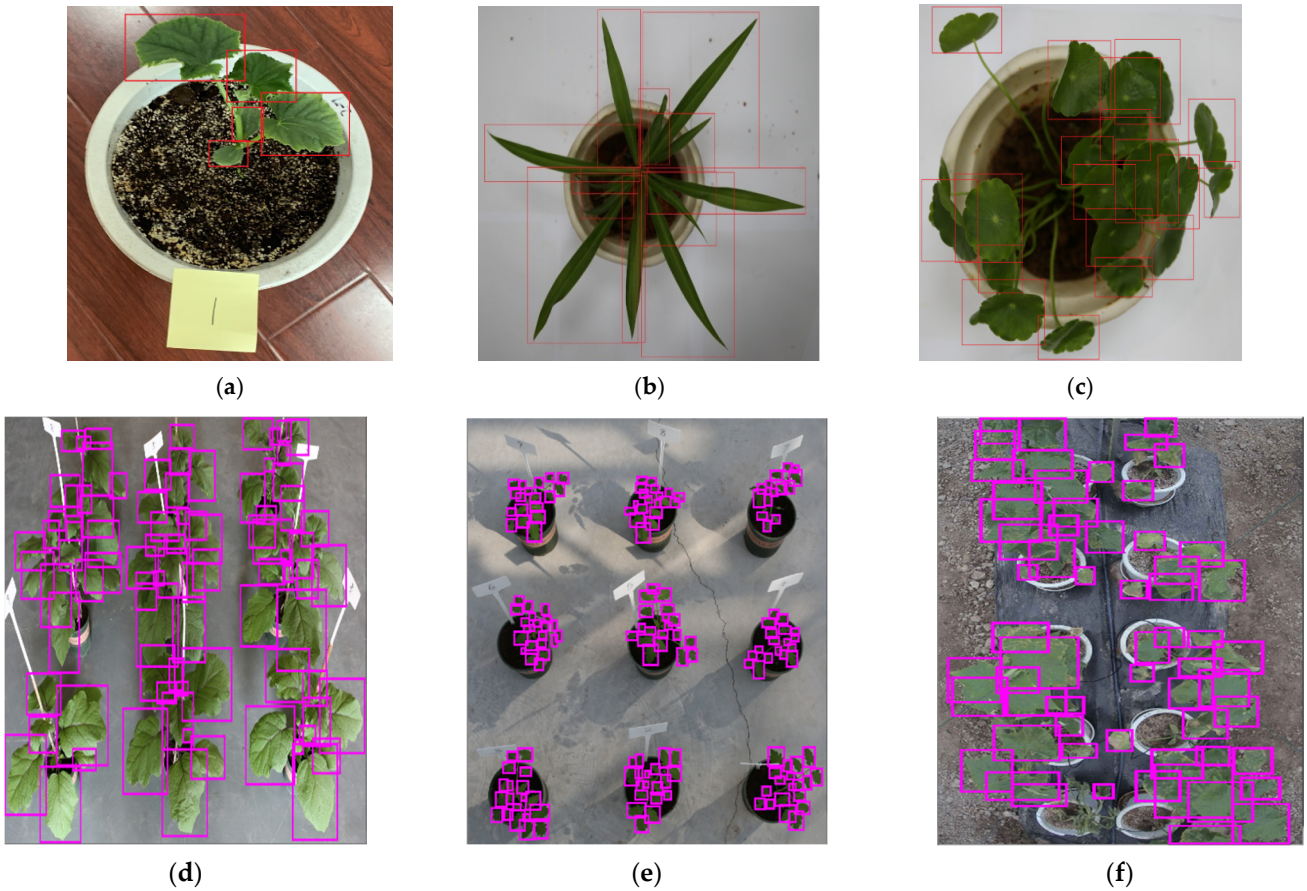

Figure 10. Visual detected results on different plants; (**a**) individual cucumber; (**b**) orchid grass; (**c**) pennywort; (**d**) eggplants; (**e**) tomato; (**f**) group cucumber.

3.5. Detection Performance for Leaves under Different Light Condition and Background

For images collected under different light conditions, the background and leaf color/texture usually have significant influence on leaf detection. Therefore, light condi-

tion and background are usually the most critical issues for detecting leaves in natural environments. To explore the possible influence of light condition and background, we used eggplant as a representative, and took images every 2 h between 6:00–18:00 under indoor/outdoor environments (both under sunny weather condition). These images were detected by the improved CenterNet detector to record subtle differences within these 12 h (the light condition is most sufficient at 12:00, while most uniform at 8:00). The average detection accuracy is shown in Figure 11.

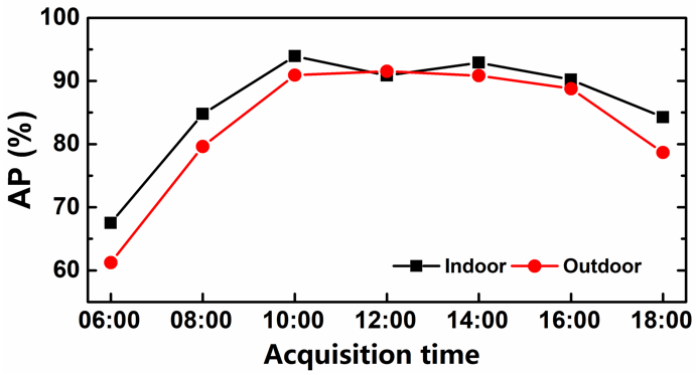

Figure 11. Average Detection Accuracy of Eggplant Leaves in Different Time Periods.

The four charts in Figure 12 showed a general satisfactory detection performance for both indoor and outdoor images, indicating that our model is not sensitive to backgrounds under sufficient light conditions. While by comparing the results with the same background, more leaves can be detected under sufficient light conditions. The robustness of our model is further demonstrated by the proper detection of many overlapped leaves, despite the influence of shadows.

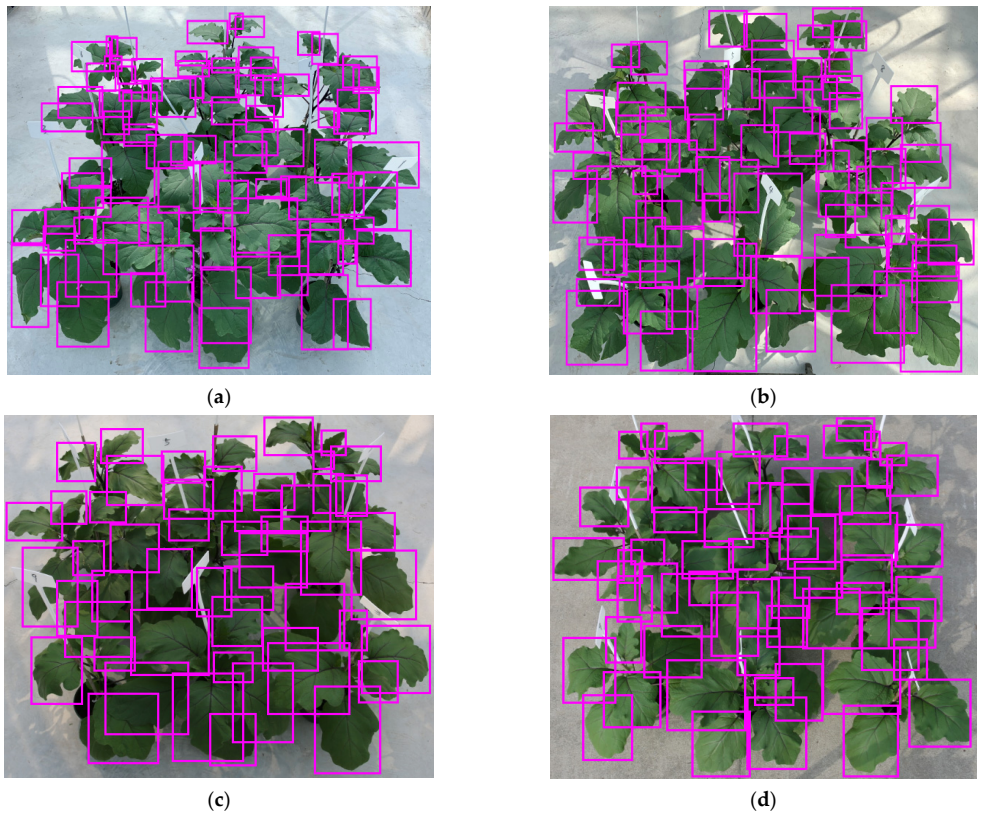

Figure 12. Detection results of eggplant leaf under different indoor and outdoor light conditions: (**a**) weak indoor light; (**b**) sufficient indoor light; (**c**) weak outdoor light; (**d**) sufficient outdoor light.

In Figure 11, the detection at moderate light condition shows higher accuracy than that at 12:00 with high-intensity direct light. Table 2 shows that leaf detections at 10:00 and 14:00 also provide relatively high accuracy with low false alarm and missed detection. Comparing the data of correct detection ratio, false alarm ratio and missed detection ratio in Table 2, the improved CenterNet shows better performance than the original CenterNet.

Table 2. Comparison of leaf detection accuracy for images acquired at different outdoor time periods. 50 images were collected for each time period, each image contains 60 leaves.

Acquisition Time	Correct Detection Ratio (%)		False Alarm Ratio (%)		Missed Detection Ratio (%)	
	Original CenterNet	Improved CenterNet	Original CenterNet	Improved CenterNet	Original CenterNet	Improved CenterNet
8:00	83.33	86.67	8.33	3.33	16.67	13.33
10:00	81.67	93.33	3.33	0	18.33	6.67
12:00	85	90	6.67	6.67	15	10
14:00	86.67	93.33	5	1.67	13.33	6.67
16:00	83.33	91.67	8.33	5	16.67	8.33
18:00	78.33	85	10	6.67	21.67	15

4. Conclusions

The segmentation-based object detection algorithms for leaf counting employed by previous studies have relatively high false rates and cannot provide clear segmentations for dense and overlapped leaves. Additional preprocessing is required due to the limitation of these algorithms in terms of specific light conditions and leaf shapes, which generates high costs and low efficiencies. Consequently, it is still a key challenge to develop an algorithm that can circumvent the overlapping problem for leaf detection and leaf counting. To address this issue, here we developed the improved CenterNet, i.e., a key-point-based one-stage object detection deep learning detector. Some specific contributions and conclusions of the improvement can be reached as follows:

(1) Our proposed improved CenterNet included three major improvements: (i) a space-to-depth module was added to the inputs of the network in the original CenterNet, for converting the input images into different depth modules; (ii) a convolutional block attention module (CBAM) was adopted to detect leaf edge information at different resolutions; and (iii) atrous spatial pyramid pooling (ASPP) was also adopted to extract image receptive field features at different scales, in case of preserving the characteristics of dense and irregular leaves;

(2) The improved CenterNet focused on the key points of the target area instead of the entire image, thereby spared the need for target segmentation and improved the accuracy and speed of detection. Apart from that, accurate detection of overlapped leaves necessitated the edge information of each leaf, along with specific constraints to suppress the background noise of the image. Traditional target detection algorithms did not make full use of the feature information of the target, as most of these algorithms used multi-scale fusion, deep network structure and loss function to obtain target features. Compare to traditional algorithms, our improved CenterNet leveraged a space-to-depth module and an attention mechanism module to enable the detection network to focus on the edge characteristics of leaves, thus facilitating the extraction of edge feature information and leaf counting.

(3) The improved CenterNet detector proposed in this work accurately detected leaves at different growth stages, under different light conditions, and of different shapes and sizes. A better detection performance was achieved, compared to the commonly used deep learning-based algorithms (including Mask R-CNN, Faster R-CNN, YOLOv4 and original CenterNet) and traditional machine learning algorithm (ExG + SVM), which had almost the highest detection speed and the minimal training time on precision and recall. The detection accuracy of the improved CenterNet detector on

images of the early growth stage was better than that on images of the late growth stage. The reason can be mostly attributed to fact that there are more leaves and occlusion in the canopy at the late growth stage of plant. Additionally, the detection accuracy of indoor images was slightly higher than that of outdoor images at different time nodes during daytime. This may result from good light conditions of the indoor images, which eliminated the influence of the bright sunlight.

Overall, this study verified the superiority of the improved CenterNet detector for detecting and counting dense and overlapped plant leaves. We believe this improved CenterNet detector will be a powerful tool in leaf detection, which is a key step in the automatic acquisition of architectural phenotypic parameters. In this study, the test was conducted only on six plant species. It is expected to verify the method on more plant species with different densities and shapes of leaves. This verification has been included in our future work. In addition, this method should be further integrated into a robotic arm with automatic control functions to continuously monitor the growth rate of plants. Another further attempt should be made to identify smaller organs such as flowers and early fruits.

Author Contributions: S.L. and T.Q. designed the experiments, provided funding, and revised the manuscript; G.L. carried out the experiments, labeled data and revised the manuscript; Z.S. wrote the manuscript; Z.S. and W.C. developed the algorithm, trained the models, and performed the analysis; M.C. provided constructive discussions; Y.Z. edited English language. All authors have read and agreed to the published version of the manuscript.

Funding: This work is supported by the National Natural Science Foundation of China (no. 61762013 and 61662006), the Science and Technology Foundation of Guangxi Province (no. 2018AD19339), Innovation Project of Guangxi Graduate Education (No. JGY2021037, XYCSZ2021007) and Research Fund of Guangxi Key Lab of Multi-source Information Mining & Security (No. 20-A-02-02).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Cobb, J.N.; DeClerck, G.; Greenberg, A. Next-generation phenotyping: Requirements and strategies for enhancing our understanding of genotype–phenotype relationships and its relevance to crop improvement. *Theor. Appl. Genet.* **2013**, *126*, 867–887. [CrossRef]
2. Rasheed, A.; Hao, Y.; Xia, X. Crop breeding chips and genotyping platforms: Progress, challenges, and perspectives. *Mol. Plant* **2017**, *10*, 1047–1064. [CrossRef] [PubMed]
3. Yang, W.; Feng, H.; Zhang, X. Crop phenomics and high-throughput phenotyping: Past decades, current challenges, and future perspectives. *Mol. Plant* **2020**, *13*, 187–214. [CrossRef]
4. Bao, Y.; Tang, L.; Breitzman, M.W. Field-based robotic phenotyping of sorghum plant architecture using stereo vision. *J. Field Robot.* **2018**, *36*, 397–415. [CrossRef]
5. Achanta, R.; Shaji, A.; Smith, K. SLIC superpixels compared to state-of-the-art superpixel methods. *IEEE Trans. Pattern Anal. Mach. Intell.* **2012**, *34*, 2274–2282. [CrossRef] [PubMed]
6. Tang, X.; Liu, M.; Zhao, H. Leaf extraction from complicated background. In Proceedings of the 2nd International Congress on Image and Signal Processing, Tianjin, China, 17–19 October 2009; pp. 1–5.
7. Yin, X.; Liu, X.; Chen, J. Multi-leaf tracking from fluorescence plant videos. In Proceedings of the IEEE International Conference on Image Processing (ICIP), Paris, France, 27–30 October 2014; pp. 408–412.
8. Vylder, J.D.; Ochoa, D.; Philips, W. Leaf segmentation and tracking using probabilistic parametric active contours. In Proceedings of the International Conference on Computer Vision/Computer Graphics Collaboration Techniques and Applications, Rocquencourt, France, 10–11 October 2011; pp. 75–85.
9. Dellen, B.; Scharr, H.; Torras, C. Growth signatures of rosette plants from time-lapse video. *IEEE/ACM Trans. Comput. Biol. Bioinform.* **2015**, *12*, 1470–1478. [CrossRef] [PubMed]

10. Grand-Brochier, M.; Vacavant, A.; Cerutti, G. Tree leaves extraction in natural images: Comparative study of preprocessing tools and segmentation methods. *IEEE Trans. Image Process* **2015**, *24*, 1549–1560. [CrossRef] [PubMed]
11. Cerutti, G.; Tougne, L.; Vacavant, A. A parametric active polygon for leaf segmentation and shape estimation. *Int. Symp. Vis. Comput.* **2011**, *6938*, 202–213.
12. Pape, J.-M.; Klukas, C. 3-D histogram-based segmentation and leaf detection for rosette plants. In Proceedings of the European Conference on Computer Vision, Zurich, Switzerland, 5–12 September 2014; pp. 61–74.
13. Scharr, H.; Minervini, M.; French, A.P. Leaf segmentation in plant phenotyping: A collation study. *Mach. Vis. Appl.* **2016**, *27*, 585–606. [CrossRef]
14. Vukadinovic, D.; Polder, G. Watershed and supervised classification based fully automated method for separate leaf segmentation. In Proceedings of the The Netherlands Congress on Computer Vision, Lunteren, The Netherlands, 14–15 September 2015; pp. 1–2.
15. Giuffrida, M.V.; Minervini, M.; Tsaftaris, S.A. Learning to count leaves in rosette plants. In Proceedings of the Computer Vision Problems in Plant Phenotyping (CVPPP), Swansea, UK, 7–10 September 2015; pp. 1–13.
16. Arteta, C.; Lempitsky, V.; Noble, J.A. Interactive Object Counting. In Proceedings of the European Conference on Computer Vision, Zurich, Switzerland, 6–12 September 2014; pp. 504–518.
17. Fiaschi, L.; Nair, R.; Koethe, U. Learning to Count with Regression Forest and Structured Labels. In Proceedings of the 21st International Conference on Pattern Recognition (ICPR 2012), Tsukuba, Japan, 11–15 November 2012; pp. 2685–2688.
18. Minervini, M.; Scharr, H.; Tsaftaris, S.A. Image analysis: The new bottleneck in plant phenotyping. *IEEE Signal Process. Mag.* **2015**, *32*, 126–131. [CrossRef]
19. Jiang, Y.; Li, C. Convolutional neural networks for image-based high-throughput plant phenotyping: A review. *Plant Phenomics* **2020**, *2020*, 4152816. [CrossRef] [PubMed]
20. Buzzy, M.; Thesma, V.; Davoodi, M. Real-time plant leaf counting using deep object detection networks. *Sensors* **2020**, *20*, 6896. [CrossRef] [PubMed]
21. Romera-Paredes, B.; Torr, P.H.S. Recurrent Instance Segmentation. In Proceedings of the Computer Vision—ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, 11–14 October 2016; pp. 312–329.
22. Donahue, J.; Hendricks, L.A.; Guadarrama, S. Long-term recurrent convolutional networks for visual recognition and description. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; Volume 39, pp. 677–691. [CrossRef]
23. Ren, M.; Zemel, R. End-to-End Instance Segmentation with Recurrent Attention. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 293–301.
24. Aich, S.; Stavness, I. Leaf counting with deep convolutional and deconvolutional networks. In Proceedings of the 2017 IEEE International Conference on Computer Vision Workshops (ICCVW), Venice, Italy, 22–29 October 2017; pp. 22–29.
25. Giuffrida, M.V.; Doerner, P.; Tsaftaris, S.A. Pheno-deep counter: A unified and versatile deep learning architecture for leaf counting. *Plant J.* **2018**, *96*, 880–890. [CrossRef]
26. Kumar, J.P.; Domnic, S. Rosette plant segmentation with leaf count using orthogonal transform and deep convolutional neural network. *Mach. Vis. Appl.* **2020**, *31*, 1–14.
27. Zhou, X.; Wang, D.; Krähenbühl, P. Objects as points. *arXiv* **2019**, arXiv:1904.07850.
28. Lin, T.-Y.; Maire, M.; Belongie, S. Microsoft coco: Common objects in context. In Proceedings of the European Conference on Computer vision, Zurich, Switzerland, 5–12 September 2014; pp. 740–755.
29. Redmon, J.; Divvala, S.; Girshick, R.; Farhadi, A. You only look once: Unified, real-time object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 779–788.
30. Woo, S.; Park, J.; Lee, J.-Y. Cbam: Convolutional block attention module. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 3–19.
31. Chen, L.-C.; Papandreou, G.; Kokkinos, I. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *40*, 834–848. [CrossRef] [PubMed]
32. He, K.; Zhang, X.; Ren, S. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
33. Li, G.; Xie, H.; Yan, W. Detection of Road Objects With Small Appearance in Images for Autonomous Driving in Various Traffic Situations Using a Deep Learning Based Approach. *IEEE Access* **2020**, *8*, 211164–211172. [CrossRef]
34. Chen, Y.M.; Zhou, D.W. Adaptive attention network for image super-resolution. *Acta Autom. Sin.* **2020**, *46*, 1–11.
35. Lazebnik, S.; Schmid, C.; Ponce, J. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In Proceedings of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06), New York, NY, USA, 17–22 June 2006; pp. 2169–2178.
36. Girshick, R. Fast r-cnn. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 7–13 December 2015; pp. 1440–1448.
37. Lin, T.-Y.; Goyal, P.; Girshick, R. Focal loss for dense object detection. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 2980–2988.
38. He, K.; Gkioxari, G.; Dollár, P. Mask r-cnn. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 2961–2969.

39. Ren, S.; He, K.; Girshick, R. Faster R-CNN: Towards real-time object detection with region proposal networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **2016**, *39*, 1137–1149. [CrossRef]
40. Bochkovskiy, A.; Wang, C.-Y.; Liao, H.-Y.M. Yolov4: Optimal speed and accuracy of object detection. *arXiv* **2020**, arXiv:2004.10934.
41. Yuan, W.; Wijewardane, N.K.; Jenkins, S. Early prediction of soybean traits through color and texture features of canopy RGB imagery. *Sci. Rep.* **2019**, *9*, 14089. [CrossRef] [PubMed]
42. Mehta, S.S.; Ton, C.; Asundi, S. Multiple camera fruit localization using a particle filter. *Comput. Electron. Agric.* **2017**, *142*, 139–154. [CrossRef]
43. Kingma, D.P.; Ba, J. Adam: A method for stochastic optimization. *arXiv* **2014**, arXiv:1412.6980.